Nucleic Acids Research, 2004, Vol. 32, Database issue ***D383–D387***
DOI: 10.1093/nar/gkh001

The Rice PIPELINE: a unification tool for plant functional genomics

Junshi Yazaki, Keiichi Kojima[1], Kouji Suzuki[1], Naoki Kishimoto and Shoshi Kikuchi*

Department of Molecular Biology, National Institute of Agrobiological Sciences, 2-1-2 Kannon-dai, Tsukuba, Ibaraki 305-8602, Japan and [1]Hitachi Software Engineering Co., Ltd, 4-12-7, Higashishinagawa, Shinagawa-ku, Tokyo 140-0002, Japan

Received June 18, 2003; Accepted July 23, 2003

ABSTRACT

The Rice Genome Research Project in Japan performs genome sequencing and comprehensive expression profiling, constructs genetic and physical maps, collects full-length cDNAs and generates mutant lines, all aimed at improving the breeding of the rice plant as a food source. The National Institute of Agrobiological Sciences in Tsukuba, Japan, has accumulated numerous rice biological resources and has already successfully produced a high-quality genome sequence, a high-density genetic map with 3000 markers, 30 000 full-length cDNAs, over 700 expression profiles with a 9000 cDNA microarray and 15 000 flanking sequences with Tos17 insertions in about 3765 mutant lines from about 50 000 transposon insertion lines. These resources are available in the public domain. A new unification tool for functional genomics, called Rice PIPELINE, has also been developed for the dynamic collection and compilation of genomics data (genome sequences, full-length cDNAs, gene expression profiles, mutant lines, *cis* elements) from various databases. The mission of Rice PIPELINE is to provide a unique scientific resource that pools publicly available rice genomic data for search by clone sequence, clone name, GenBank accession number, or keyword. The web-based form of Rice PIPELINE is available at http://cdna01.dna.affrc.go.jp/PIPE/.

INTRODUCTION

In the past, biologists have made great efforts to isolate useful genes one at a time by using genetic and molecular biological approaches such as positional cloning (in which the map position of a gene is determined by using various DNA markers) and by finding mutated DNA sequences on the gene. Now, in the post-genome-sequence era, the style of genomic research has changed dramatically.

In research on rice (*Oryza sativa*), the National Institute of Agrobiological Sciences (NIAS) at Tsukuba, Japan, and its collaborators have constructed useful tools for functional genomics through the Rice Genome Project of Japan. These tools consist of approximately 30 000 full-length cDNAs, over 700 expression profiles developed by using an 8987-EST microarray, a high-quality genomic sequence that has 99.99% accuracy, a genetic map with approximately 3200 DNA markers, and about 50 000 transposon insertion lines. All of this material is accessible to the public from the Knowledge-based Oryza Molecular biological Encyclopedia (1) (KOME: see URL in Table 1); the Rice Expression Database (2) (RED: see URL in Table 1); the INtegrated rice genome Explorer (3) (INE; see URL in Table 1); the *cis* element database (PLACE: see URL in Table 1), the Tos17 mutant panel database (Tos17: see URL in Table 1); and the Ministry of Agriculture, Forestry and Fisheries DNA Bank (MAFF DNA Bank: see URL in Table 1) at NIAS (see URL in Table 1). These tools greatly increase the efficiency and accuracy of research on rice functional genomics, such as the identification of the genes that control particular phenotypes.

However, we still need a unification tool that can be used dynamically to collect and collate genomics data from many databases on the basis of the input of just one sequence, gene ID, or keyword.

Such unification tools have been developed for mouse and worm genomics and are capable of gathering genomics tools and information world-wide. Integrated genomics tools are publicly accessible at Mouse Genome Informatics (www.informatics.jax.org/), FANTOM (4) (fantom2.gsc.riken.go.jp/) and WormBase (5) (www.wormbase.org/).

In the plant world, biological tools for functional genomics already exist. For example, for *Arabidopsis thaliana*, we have the *Arabidopsis* Information Resource (TAIR: arabidopsis.org/home.html); the *Arabidopsis* EST Index (6) (www.kazusa.or.jp/en/plant/arabi/EST/); and the RIKEN *Arabidopsis* full-length cDNA database (7) (pfgweb.gsc.riken.go.jp/pub_data/full_length.html). However, the genomics resources for such databases are scattered globally. In contrast, the biological resources for rice research in Japan have been gathered centrally by NIAS and its collaborators. We therefore have been able to develop a search unification tool, Rice PIPELINE, that can be used to find various types of genomics information, for example, to identify a gene or to find a mutated DNA sequence on the gene.

Here, we describe how the tool can be used for computer-based study of rice functional genomics.

*To whom correspondence should be addressed. Tel/Fax: +81 298 387007; Email: skikuchi@nias.affrc.go.jp

Table 1. Useful URLs for rice genomics and sites related to Rice PIPELINE

NIAS homepage	www.nias.affrc.go.jp/index_e.html
Unified tool <Rice PIPELINE>	cdna01.dna.affrc.go.jp/PIPE/
Genome map & sequence <INE>	rgp.dna.affrc.go.jp/giot/INE.html
Full length cDNA <KOME>	cdna01.dna.affrc.go.jp/cDNA/
Expression profile <RED>	red.dna.affrc.go.jp/RED/
Mutant line <Tos17>	tos.nias.affrc.go.jp/
Plant *cis* element <PLACE>	www.dna.affrc.go.jp/htdocs/PLACE/
Genome resource <MAFF DNA Bank>	www.dna.affrc.go.jp/
Rice Microarray Opening Site <RMOS>	microarray.rice.dna.affrc.go.jp
Genome consortium <IRGSP>	rgp.dna.affrc.go.jp/IRGSP/
Genome research program <RGP>	rgp.dna.affrc.go.jp/
Genome database <Oryzabase>	www.shigen.nig.ac.jp/rice/oryzabase/
Genome annotation system <RiceGAAS>	RiceGAAS.dna.affrc.go.jp/
Genome annotation <RiceHMM>	rgp.dna.affrc.go.jp/RiceHMM/

FEATURES OF THE TOOL

Rice PIPELINE is a unification tool that dynamically collects and collates data from various databases (KOME, INE, RED, Tos17, PLACE; see URLs in Table 1) at NIAS, and thereby attempts to present the genetics and molecular biology of *O.sativa* in an easy-to-navigate site. The mission of Rice PIPELINE (see URL in Table 1) is to provide a unique scientific resource on rice that pools publicly available data and makes them readily available through the input of any clone sequence, clone name, GenBank accession number, or keyword. For example, in the case of a sequence query, users can select the BLAST condition, the type of full-length cDNA information (nucleic acid, amino acid), *indica* and *japonica* genome information, and the microarray EST before they submit the query. The user is presented with a multiple BLAST search result (Fig. 1) that presents data from the selected database according to similarity alignments. The data are composed of microarray EST information, annotations on the nucleotides and amino acids from full-length cDNAs, and genome information from the two rice cultivars (*indica* and *japonica*). For additional information, the user can click on the clone name in 'Full-length cDNA BLASTN Result' or in 'Full-length cDNA Longest ORF BLASTX Result' to obtain the results of a *cis*-element search using PLACE (8) (see URL in Table 1), a gene expression profile, phenotype information, or other details such as domain search results and gene ontology (GO) classification. Users can also obtain expression profiles directly from multiple BLAST search results by using the 'search RED' function.

INFORMATION FLOWCHART

Figure 2 is an information flowchart showing how Rice PIPELINE is constructed. The top pathway in the figure shows the flow of gene structural information that results from a query. From Rice PIPELINE, users can obtain a GenBank report and a KOME report containing nucleic acid and amino acid analyses, domain search results, and GO information. Via the KOME report, they can also obtain further information on the *cis*-element motif at the 5′ upstream region of the full-length cDNA by a PLACE search, phenotype information with flanking sequences from Tos17, and an expression profile from RED. Such information can be used for isolating transcriptional factors by computer and for elucidating gene structure.

The middle pathway in Figure 2 shows the flow of gene expression information that results from a query. From Rice PIPELINE, users can obtain a GenBank report, a MAFF DNA Bank report, and an expression profile link with RED. Additionally, from the RED expression profile they can obtain genetic and physical map information on the INE link. The information obtained by this middle flow pathway facilitates the identification of function from expression profiling of the gene under various physiological conditions.

The bottom pathway in Figure 2 shows the flow of genome information from the BLAST results of a query. Users receive genetic and physical map information, including genome sequences, on INE as a result of a *japonica* genome search on Rice PIPELINE. The INE map information is directly linked with the GenBank report. Users can also obtain a GenBank report from an *indica* genome search. The information obtained by this bottom flow pathway facilitates high-resolution genetic mapping, positional cloning of target genes, and genetic dissection of quantitative trait loci (QTLs).

INTEGRATED DATABASES IN RICE PIPELINE

Rice PIPELINE integrates a number of databases (KOME, INE, RED, Tos17, PLACE) to make *O.sativa* data easy to navigate. It is designed to provide easily accessible and comprehensive information on rice functional genomics. Below, we outline the information provided by each database.

INE is a database that integrates the genetic map, physical map and sequencing information of the rice genome. Integrated maps are presented for each chromosome. A marker search leads directly to the data available on specific DNA clones. This information can be used to identify gene function and elucidate chromosome structure.

KOME is a database of rice full-length cDNAs of 28 469 unique genes at present (1). For all cDNAs, the user can perform full sequencing, nucleotide analysis, amino acid analysis, GO classification and digital mapping on the genome sequences of *indica* and *japonica* cultivars. The database also has *cis*-element information on each clone from PLACE, phenotype information linked with the mutant panel database Tos17 and expression information linked with RED.

PLACE is a database of motifs found in plant *cis*-acting regulatory DNA elements, all from previously published reports. It covers vascular plants only. A *cis*-element search can be performed by using PLACE to excise 1000 base pairs (bp) of genomic sequence upstream from the 5′ termini of

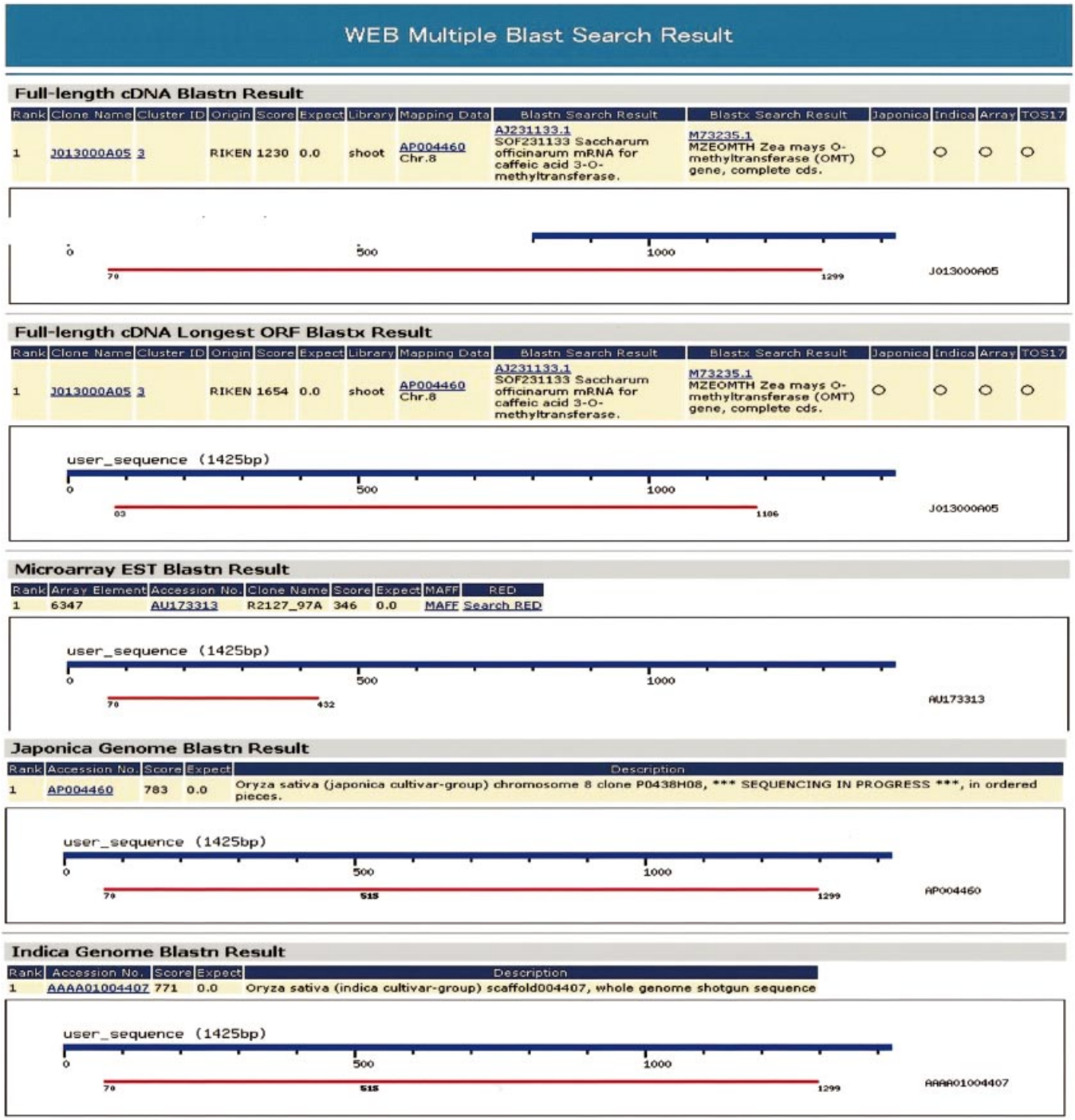

Figure 1. On-screen appearance of the results of a Rice PIPELINE search. The top two parts show the results of a query to the full-length cDNA database, KOME. The clone name is linked with KOME by the details of its full-length cDNA. The mapping data and BLASTN and BLASTX search results are directly linked with the GenBank report. The items at the right end of the result ('Japonica', 'Indica', 'Array' and 'Tos17') relate the full-length cDNA data to information on genome sequence, microarray ESTs and flanking sequence in the Tos17 mutant line. The next major item down shows the results of a microarray search of 8987 EST clones. The accession is directly linked with the GenBank report, and MAFF is directly linked with the DNA Bank Rice Database (MAFF DNA Bank: see URL in Table 1). A search of RED is also possible. The two major items at the bottom show the results of a query of the *indica* and *japonica* genome sequences. The accession corresponding to the *japonica* genome BLASTN result is linked with genetic and physical map information on the INE Database. The accession corresponding to the *indica* genome BLASTN result is linked with the GenBank report. Each alignment map shows the match between the query and result sequences. The red bar in each result sequence is linked with the KOME report, INE information, the MAFF DNA Bank report and the GenBank report.

each full-length cDNA clone and search about 300 *cis* elements known from plants. In Rice PIPELINE, the results of the *cis*-element search can be obtained from the results of a full-length cDNA BLASTN or BLASTX search as a multiple BLAST search result. Users can also obtain full-length cDNA with similarities to microarray ESTs at the Rice Microarray

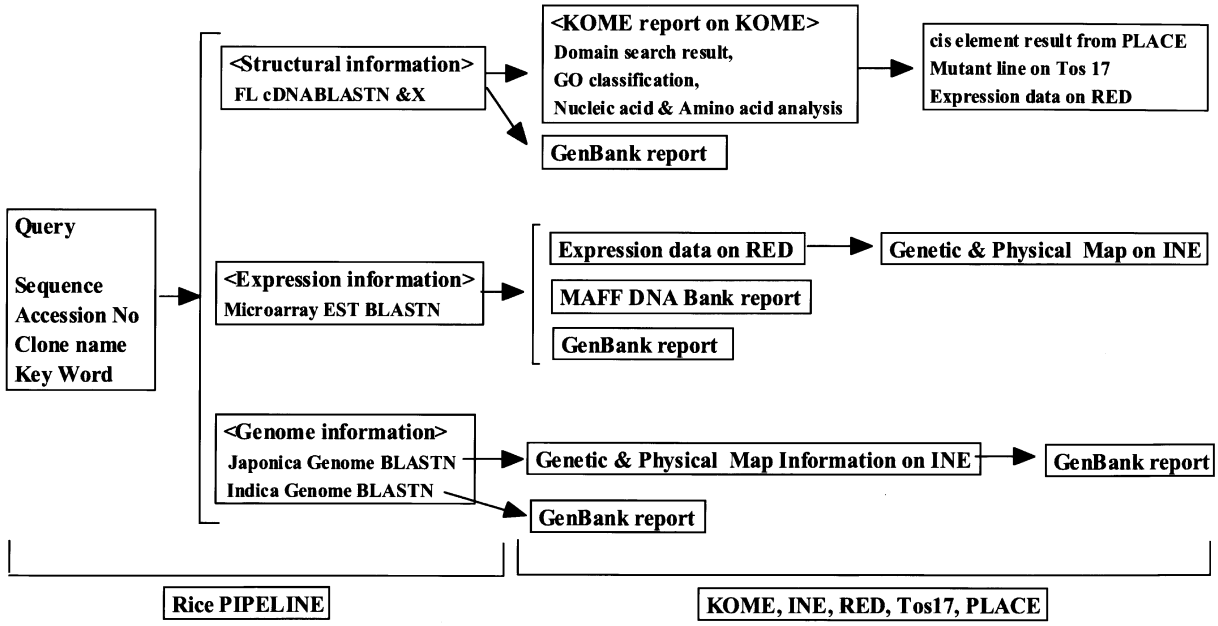

Figure 2. Flow of information in Rice PIPELINE. The top pathway shows the flow of gene structural information from the query. Users can obtain a GenBank report and a KOME report containing nucleic acid and amino acid analyses, domain search results and GO classification from Rice PIPELINE. They can also obtain information on *cis*-element motifs, phenotypes with flanking sequences, and expression profiles via the KOME report. The middle pathway shows the flow of gene expression information from the query. Users can obtain a GenBank report, a MAFF DNA Bank report, and an expression profile from Rice PIPELINE. Additionally, they can use genetic and physical map information on an INE link with RED. The bottom pathway shows the flow of genome information from BLAST results. Users can utilize map information, including genome sequences, from a PIPELINE search of the *japonica* genome. The map information of INE is directly linked with the GenBank report. Users can also obtain a GenBank report from the results of a PIPELINE search of the *indica* genome.

Opening Site (RMOS: see URL in Table 1). This function enables a search of *cis*-elements from the expression profiles of ESTs on the rice microarray.

The rice retrotransposon Tos17, which is highly activated by tissue culture, has been used for insertional mutagenesis of rice. NIAS has collected 15 111 flanking sequences with Tos17 insertions in about 3765 mutant lines from about 50 000 insertion lines. The Tos17 mutant panel database enables the user to link flanking sequences with phenotype information by using BLAST. With this database, users can perform gene function analysis by computer and reverse genetics. To search these mutant lines, users are required to register on the Tos17 site (see URL in Table 1).

RED contains normalized expression data derived from experiments using various RNAs hybridized to the Rice 9000 cDNA Array (over 700 experiments in 26 physiological categories). All of the expression data are shown as values of the expression ratio. The web-based form of RED can also be directly accessed from RMOS. RMOS includes information on our microarray system, including microarray design, EST clone information linked with the full-length cDNA information of KOME, and our experimental system and protocols.

The Rice Genome Resource Center (http://www.rgrc.dna.affrc.go.jp/index.html.en) will make the above material available world-wide from 2004.

THE FUTURE

We are already beginning to integrate rice genomics information accumulated by NIAS and phenotype-level information on various lines and mutants with genetic map information in the Oryzabase (see URL in Table 1) run by the National Institute of Genetics. This integration will make it possible for scientists to view information at both the molecular and phenotype levels.

RICE GENOMICS IN JAPAN

The rice genome sequence was determined at a highly accurate level (99.99%) by the efforts of the International Rice Genome Sequencing Project (9,10), Syngenta (11) and the Beijing Genomics Institute (12). We now have the whole genome sequence of a monocot plant (*O.sativa*) and a dicot (*A.thaliana*).

Because of the availability of this high-quality sequence information, the research stage will now shift from genome sequencing to functional genomics. Rice is not only a very important agricultural resource but also a model plant for biological research. There are now many kinds of genomics tools for rice, including high-quality sequence data, full-length cDNAs, ESTs (13), DNA chip technology (14), information on mutant lines, and genetic (15) and physical maps (16). Rice and *A.thaliana* will continue to be used in research because of their importance in the genetic improvement of crops, although in Japan the resources available for rice research are greater than those for *Arabidopsis*. Genes for controlling heading date (17), growth and development (18), vivipary-related genes (19), and genes involved in adaptation to Fe-deficiency (20) and control of flour quality (H.Shimada, Science University of Tokyo, personal communication) have

already been isolated from *O.sativa* at NIAS by using various functional genomics tools and materials. This rice research will continue to help unlock the genetic secrets of plants and develop rice as a better food resource world-wide.

ACKNOWLEDGEMENTS

We thank Yuko Nagata, Akiko Hashimoto, Kanako Shimbo, Yumiko Yoshida, Zenpei Shimatani (STAFF-Institute), Dr Masahiro Ishikawa, Fumiko Fujii, Keiko Takeuchi, Kazuko Toyoshima, Yuki Sato, Chikako Miyamoto, Sachiko Honda, Ayano Endo (National Institute of Agrobiological Sciences), for helpful support, Yoshiyuki Mukai (STAFF-Institute) and Makoto Yamamoto (Hitachi Software Engineering) for useful advice. This work was supported by a grant from the Ministry of Agriculture, Forestry, and Fisheries of Japan (Rice Genome Project SY-1112).

REFERENCES

1. Kikuchi,S., Satoh,K., Nagata,T., Kawagashira,N., Doi,K., Kishimoto,N., Yazaki,J., Ishikawa,M., Ooka,H., Kojima,K. *et al.* (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science*, **301**, 376–379.
2. Yazaki,J., Kishimoto,N., Ishikawa,M. and Kikuchi,S. (2002) Rice Expression Database: the gateway to rice functional genomics. *Trends Plant Sci.*, **7**, 563–564.
3. Sakata,K., Antonio,B.A., Mukai,Y., Nagasaki,H., Sakai,Y., Makino,K. and Sasaki,T. (2000) INE: a rice genome database with an integrated map view. *Nucleic Acids Res.*, **28**, 97–102.
4. Okazaki,Y., Furuno,M., Kasukawa,T., Adachi,J., Bono,H., Kondo,S., Nikaido,I., Osato,N., Saito,R. and Suzuki,H. (2002) Analysis of the mouse transcriptome based on functional annotation of 60 770 full-length cDNAs. *Nature*, **420**, 563–573.
5. Harris,T.W., Lee,R., Schwarz,E., Bradnam,K., Lawson,D., Chen,W., Blasier,D., Kenny,E., Cunningham,F., Kishore,R. *et al.* (2003) WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res.*, **31**, 133–137.
6. Asamizu,E., Nakamura,Y., Sato,S. and Tabata,S. (2000) A large scale analysis of cDNA in *Arabidopsis thaliana*: Generation of 12 028 non-redundant expressed sequence tags from normalized and size-selected cDNA libraries. *DNA Res.*, **7**, 175–180.
7. Seki,M., Narusaka,M., Kamiya,A., Ishida,J., Satou,M., Sakurai,T., Nakajima,M., Enju,A., Akiyama,K., Oono,Y. *et al.* (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science*, **296**, 141–145
8. Higo,K., Ugawa,Y., Iwamoto,M. and Korenaga,T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Res.*, **27**, 297–300.
9. Sasaki,T., Matsumoto,T., Yamamoto,K., Sakata,K., Baba,T., Katayose,Y., Wu,J., Niimura,Y., Cheng,Z., Nagamura,Y. *et al.* (2002) The genome sequence and structure of rice chromosome 1. *Nature*, **420**, 312–316.
10. Feng,Q., Zhang,Y., Hao,P., Wang,S., Fu,G., Huang,Y., Li,Y., Zhu,J., Liu,Y., Hu,X. *et al.* (2002) Sequence and analysis of rice chromosome 4. *Nature*, **420**, 316–320.
11. Goff,S.A., Ricke,D., Lan,T.H., Presting,G., Wang,R., Dunn,M., Glazebrook,J., Sessions,A., Oeller,P., Varma,H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, 92–100.
12. Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.
13. Yamamoto,K. and Sasaki,T. (1997) Large-scale EST sequencing in rice. *Plant Mol. Biol.*, **35**, 135–144.
14. Yazaki,J., Kishimoto,N., Nakamura,K., Fujii,F., Shimbo,K., Otsuka,Y., Wu,J., Yamamoto,K., Sakata,K., Sasaki,T. *et al.* (2000) Embarking on rice functional genomics via cDNA microarray: use of 3′ UTR probes for specific gene expression analysis. *DNA Res.*, **7**, 367–370.
15. Harushima,Y., Yano,M., Shomura,A., Sato,M., Shimano,T., Kuboki,Y., Yamamoto,T., Lin,S.Y., Antonio,B.A., Parco,A. *et al.* (1998) A high-density rice genetic linkage map with 2275 markers using a single F2 population. *Genetics*, **148**, 479–494.
16. Wu,J., Maehara,T., Shimokawa,T., Yamamoto,S., Harada,C., Takazaki,Y., Ono,N., Mukai,Y., Koike,K., Yazaki,J. *et al.* (2002) A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell*, **14**, 525–535.
17. Yano,M. (2001) Genetic and molecular dissection of naturally occurring variation. *Curr. Opin. Plant Biol.*, **4**, 130–135.
18. Ashikari,M., Wu,J., Yano,M., Sasaki,T. and Yoshimura,A. (1999) Rice gibberellin-insensitive dwarf mutant gene Dwarf 1 encodes the alpha-subunit of GTP-binding protein. *Proc. Natl Acad. Sci. USA*, **96**, 10284–10289.
19. Agrawal,G.K., Yamazaki,M., Kobayashi,M., Hirochika,R., Miyao,A. and Hirochika,H. (2001) Screening of the rice viviparous mutants generated by endogenous retrotransposon Tos17 insertion. Tagging of a zeaxanthin epoxidase gene and a novel ostatc gene. *Plant Physiol.*, **125**, 1248–1257.
20. Negishi,T., Nakanishi,H., Yazaki,J., Kishimoto,N., Fujii,F., Shimbo,K., Yamamoto,K., Sakata,K., Sasaki,T., Kikuchi,S. *et al.* (2002) cDNA microarray analysis of gene expression during Fe-deficiency stress in barley suggests that polar transport of vesicles is implicated in phytosiderophore secretion in Fe-deficient barley roots. *Plant J.*, **30**, 83–94.